An Automated Framework for Advertisement Detection and Removal from Sports Videos using Audio-Visual Cues

Article *in* Frontiers of Computer Science (electronic) · December 2019

DOI: 10.1007/s11704-019-9187-9

CITATIONS
4

READS
1,325

6 authors, including:

Abeer Toheed
University of Engineering and Technology Taxila
2 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Ali Javed
University of Engineering and Technology Taxila
186 PUBLICATIONS **4,625** CITATIONS
SEE PROFILE

Aun Irtaza
University of Engineering and Technology Taxila
121 PUBLICATIONS **3,791** CITATIONS
SEE PROFILE

Hassan Dawood
University of Engineering and Technology Taxila
64 PUBLICATIONS **1,131** CITATIONS
SEE PROFILE

Front. Comput. Sci., 2021, 15(2): 152313
https://doi.org/10.1007/s11704-019-9187-9

LETTER

An automated framework for advertisement detection and removal from sports videos using audio-visual cues

Abeer TOHEED[1], Ali JAVED (✉)[1], Aun IRTAZA[1], Hassan DAWOOD[1], Hussain DAWOOD[2], Ahmed S. ALFAKEEH[3]

1 Department of Software Engineering, University of Engineering and Technology-Taxila, Taxila 47050, Pakistan
2 Department of Computer & Network Engineering, University of Jeddah, Jeddah 21577, Saudi Arabia
3 Information System Department, King Abdulaziz University, Jeddah 21577, Saudi Arabia

1 Introduction

Advertisements detection and replacement with different ads based on the user preferences is employed during sports rebroadcasts that offers more value to both the distributor and viewer. Manual advertisements detection is a laborious activity and demands an urgent need to develop automated advertisement detection techniques to save the time, storage space,and transmission bandwidth.

Existing methods use audio features [1], visual features [2,3], or hybrid features [4,5] for advertisement detection. For example, Ramires et al. [1] employed the approach of audio silence detection between shots for defining boundaries between the advertisement and program shots. Visual features usually offer better performance over audio features but at the expense of increased computational cost. Luo et al. [2] designed a deep learning based optimization framework for advertisement detection considering the global image information. Hossari et al. [3] employed the VGG19 deep neural network for advertisements detection. Existing works [4,5] also used fusion of audio-visual features to improve the accuracy of advertisements detection methods. Qian [4] proposed an advertisement detection method based on audio-visual features using the color histogram and short-time energy (STE).

Existing advertisement detection techniques have certain limitations. For example, STE audio feature-based methods are unable to accurately classify the game and advertisement shots as both contain the speech content. Logo detection-based techniques face certain challenges i.e., selection of robust distance function, position invariance, variations in image, template size, etc. Additionally, logo detection-based techniques are computationally expensivedue to template matching against various logo templates. Moreover, the techniques using local features [6] (e.g., SIFT, LBP, etc.) for shot boundary detection are sensitive to camera variations (i.e., zooming). In this letter, we propose an audio-visual features-based advertisements detection method to address the aforementioned limitations. The main contributions of the proposed framework are as follows:

- We present an effective advertisement detection framework using audio-visual features.
- Our deep learning-based shot boundary detection is robust to aforementioned limitations, i.e., variations in camera, position, image and template sizes, etc.
- Our advertisements detection method effectively captures the attributes of musical and speech tones through robust spectral features that can reliably be used to classify the advertisement and game shots.

Received March 12, 2019; accepted July 21, 2019

E-mail: ali.javed@uettaxila.edu.pk

2 Proposed method

This letter presents a two-step method for advertisements detection from the sports videos. In the first step, AlexNet deep learning model is employed for shot boundary detection to filter the candidate advertisement shots. In the second step, corresponding audio stream of these candidate shots is analyzed to further improve the accuracy by filtering out the game shots from the candidate advertisement shots. More specifically, audio signal is represented through fusion of gammatone cepstral coefficients (GTCC) and Mel-Frequency cepstral coefficients (MFCC) features. This fused features-set is then used to train the weighted K-Nearest Neighbours classifier for advertisements detection. Finally, advertisement shots are removed from the entire video to provide either advertisement free sports video. The flow of the proposed framework is shown in Fig. 1.

2.1 Shot boundary detection using AlexNet CNN

In the proposed work, we employ AlexNet convolutional neural network (CNN) for shot boundary detection. The proposed AlexNet CNN framework comprises of 25 layers, where we used five convolutional layers, two fully connected hidden layers and one fully connected output layer. Moreover, we used three maximum pooling layers of size 3×3 and stride of 2 after the first, second and fifth convolutional layer. We employed seven rectified linear unit (ReLU) layers where one ReLU layer is placed after every convolutional layer and first two fully connected layers. Additionally, two Cross channel normalization

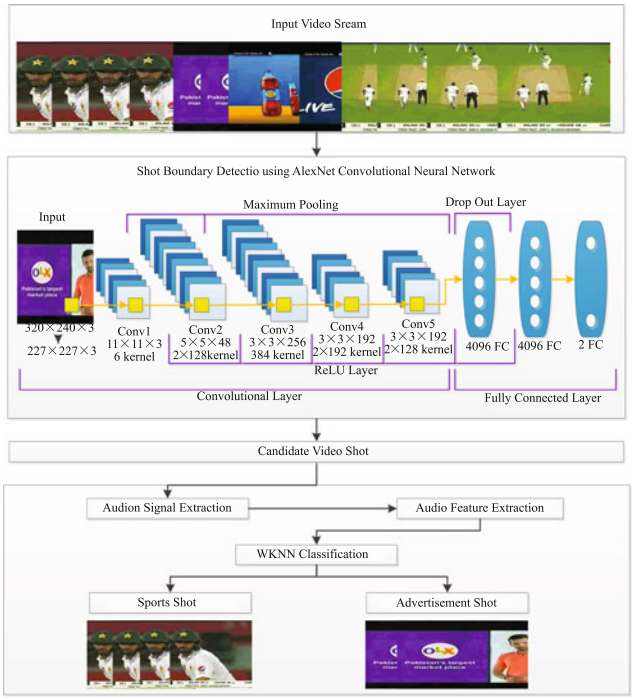

Fig. 1 Block diagram of proposed framework

layers with window size of 5 are also employed before the second and third convolutional layers. We used two dropout layers with dropout value of 0.5 before the first two fully connected layers, and one softmax layer is used after the last fully connected layer. The classification output layer is the last layer of our proposed architecture. Our first convolutional layer is image input layer with size of 227×227×3. Since, our video dataset has frame resolution of 320×240×3, therefore we down-sample our frame resolution to 227×227×3 to support the AlexNet framework. The candidate frame is convolved with first convolutional layer of our deep learner by applying the filter/window of 11×11. The window size in the second convolutional layer is reduced to 5×5, followed by 3×3 in the third, fourth and fifth convolutional layer. We used different number of kernels at each convolutional layer, i.e., 96 kernels at first convolutional layer, two groups of 128 kernels at second convolutional layer, 384 kernels at third convolutional layer, two groups of 192 kernels at fourth layer and two groups of 128 kernels at fifth convolutional layer. After the fifth convolutional layer, we used two fully connected layers with 4096 nodes. The last fully connected layer comprises of two nodes is used to represent the two classes (i.e., advertisement, sports). We used the initial learning rate of 0.001, minimum batch size of 75 and 20 number of Epochs to train our network.

2.2 Audio stream analysis for advertisements detection

After employing shot boundary detection, we obtain the candidate advertisement shots. This phase of our method analyses the audio stream of the candidate advertisement shots to remove any false positives selected at the first step. After watching massive number of sports broadcasts, we observed that the advertisements contain musical notes as compared to sports videos where we experience only the speech segments. Some exceptions do exist but on a small extent. Therefore, we exploit this observation and filter the advertisement shots based on detecting musical tones in the audio streams.

2.2.1 Problem formulation

Let $y[n]$ be the audio signal having N samples associated with the input video containing K video frames represented as $I^{(i)}(x, y)$. And $y_s[n]$ be the audio signal with N' audio frames associated with the input video shot containing K' video frames denoted as $I'^{(i)}(x, y)$ where $N' \ll N$ and $K' \ll K$. For advertisement detection, we analyze audio frames of the candidate video shots and represent them as feature vectors.

2.2.2 Features extraction

Effective features representation of the input audio is required to achieve better classification performance. Since our audio processing module is based on differentiating between the music and speech segments to identify the advertisement shots. Therefore, we analyze various traits of the speech and music segments for features selection. We observe horizontal line patterns in the music spectrograms. These spectral attributes of the signal can effectively be captured using spectral audio features.

In this letter, we used MFCC and GTCC spectral features for audio representation. MFCC takes human perception sensitivity with respect to frequencies into consideration, whereas, GTCC emulates human hearing by simulating the impulse response of the auditory nerve fibre. Additionally, GTCC is more robust to noise over MFCC. To better capture the spectral attributes of the music and speech segments, we employ the fusion of GTCC and MFCC features. More specifically, we extract 26-D features consisting of 13-D MFCC and 13-D GTCC from the audio signal.

2.3 Weighted KNN classification

For classification of the advertisement and game shots, we used the MFCC-GTCC features-set to train the weighted K-Nearest Neighbor (WKNN) classifier. We employed ten-fold cross-validation scheme for training purpose and evaluated the classifier performance on the audio samples belonging to game and advertisement shots. We tuned three parameters i.e., distance weight, distance metric and number of neighbors for classifier training. We evaluated the performance of the KNN using different parameters and achieved best results on the WKNN. More specifically, we set the distance metric to Euclidean, distance weight to squared inverse, and number of neighbors to 10 for WKNN. This squared inverse scheme assigns more weightage to nearby neighbors and vice versa.

3 Experiments and results

3.1 Dataset

Performance of the proposed framework is evaluated on 35 broadcasted sports videos available on YouTube. The dataset videos are comprised of five renowned broadcasters namely *Ten Sports*, *Star Sports*, *ESPN*, *PTV Sports*, and *Geo Super*. Moreover, dataset videos belong to three sports categories that are *Soccer*, *Tennis*, and *Cricket*. We have used the YouTube videos for performance evaluation as done by the comparative methods. For this purpose, we selected the youTube videos that are diverse in illumination conditions, shots and video length,

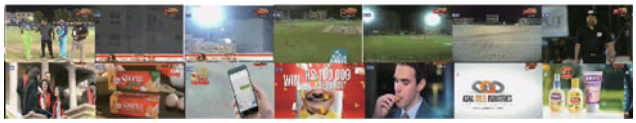

Fig. 2 Snapshot of our dataset

advertisement types, advertisements length, editing effects, etc. Each video in the dataset has a frame rate of 20 fps and a resolution of 320×240. Additionally, audio stream of our dataset videos has a sampling frequency of 44.1 KHz with 2 channels and 16 bits per sample. Shown in Fig. 2 are few snapshots of both classes of our dataset.

3.2 Performance evaluation

Performance of the proposed method is evaluated for each video in the dataset. We employed precision, recall, f1-score, accuracy, and error rate metrics to evaluate the proposed method.

In our first experiment, we computed the performance of our shot boundary detectionmethod.We selected 70% samples for training and 30% for testing the AlexNet framework. For network training, we down-sampled frames resolution to 227×227×3. We tuned following parameters during training i.e., epochs, mini-batch size, and initial learning rate. After extensive experimentation we obtained best results with initial learning rate of 0.001, mini-batch size of 75, and epochs of 20. We presented the results of five selected videos in Table 1. From the results, we can observe that the proposed framework provides remarkable results and achieves an average accuracy of 98.74%.

In our second experiment, we compared the performance of proposed AlexNet deep learning model against two renowned models i.e., GoogleNet and SqueezeNet. We employed the similar settings as adopted for AlexNet deep learning model in this experiment. Our AlexNet model comprises of 25 layers, whereas, squeezeNet and GoogleNet models use 68 and 144 layers respectively. Additionally, AlexNet deep learning model is computationally more efficient as compared to squeezeNet and GoogleNet. We experienced minimum training time of 7 mins for AlexNet, 12 mins for squeezeNet and 33 mins for GoogleNet. For performance comparison in terms of efficiency, we evaluated the performance of these three deep learning models on different dataset sizes and found the same time complexity pattern for these frameworks in all cases. We employed 70-30 holdout validation scheme for network training. Shown in Table 2 are the detection performance of these deep learning models. From the results, we can clearly observe that our AlexNet CNN framework outperforms both squeezeNet and GoogleNet models for shot boundary detection.

In our third experiment, we evaluated the performance of our audio stream analysis method for advertisement detection using MFCC, GTCC, and fusion of MFCC-GTCC features. GTCC spectral features performs slightly better than the MFCC features, however fusion of the GTCC and MFCC provides superior classification performance over using MFCC or GTCC alone. More specifically, we achieved the precision of 99.33%, recall of 99.33%, F1-score of 99.33%, accuracy of 99.37%, and error rate of 0.63% on MFCC-GTCC features fusion as shown in Table 3. Therefore, we employed the fusion of MFCC-GTCC features to train the weighted KNN for classification.

In our last experiment, we compared the performance of the proposed method against existing methods to indicate the effectiveness of our method for advertisement detection. From the results (Table 4), we can observe that our method provides superior performance over comparative approaches.

Table 1 Performance of AlexNet CNN for shot boundary detection

Video	Filtered Shots	Precision	Recall	F1-Score	Accuracy	Error
Vid1	804	100	97.39	98.68	98.68	1.32
Vid2	414	99.33	97.39	98.35	98.33	1.64
Vid3	284	98	99.32	98.66	98.67	1.33
Vid4	645	96.67	97.32	96.99	97	3
Vid5	677	98.67	97.37	98.01	98	2
Average		**98.53**	**97.76**	**98.14**	**98.74**	**1.86**

Table 2 Performance comparison of AlexNet, SqueezeNet and GoogleNet

Parameters	AlexNet	SqueezeNet	GoogleNet
Accuracy	98.67	97	97.67
Precision	98	96.67	98.67
Recall	99.32	97.32	96.73
F1-Score	98.66	96.99	97.69
Error	1.33	3	2.33

Table 3 Performance comparison of different audio spectral features

Features	Precision	Accuracy	F1_Score	Recall	Error
MFCC	88	91	90	88	9
GTCC	94	96	95	94	4
MFCC-GTCC	99	99.33	99	99	1

Table 4 Performance comparison

Advertisement Detection Methods	Features			Methodology	Accuracy
	Audio	Visual	Hybrid		
Ramires et al. [1]	X			Linear Regression	87.4%
Hossari et al. [3]		X		VGG19	94%
Zhang and Xu [5]			X	SVM-DP	94.73%
Proposed Method			**X**	**AlexNetand WKNN**	**99.37 %**

4 Conclusion

This letter presents an effective two-step advertisement detection framework using audio-visual features. We employed AlexNet deep learning model for shot boundary detection to filter the candidate advertisement shots. Next, we represented the audio of the selected video shots through MFCC-GTCC spectral features-set to train the WKNN for advertisements detection and removal from the broadcasts. The average accuracy of 99.37% signify the effectiveness of the proposed framework. Performance of our method degrades to some extent under the conditions where audio tone of the advertisement and game shots become similar or if the advertisement contains no music in the background and sports contain musical tone. We are planning to address these limitations in the future.

Acknowledgements This work was supported and funded by the Directorate ASR&TD of UET-Taxila (UET/ASR&TD/RG-1002).

References

1. Ramires A, Cocharro D, Davies M E P. An audio-only method for adver-

tisement detection in broadcast television content. 2018, arXiv preprint arXiv:1811.02411

2. Luo C, Peng Y, Zhu T, Li L. An optimization framework of video advertising: using deep learning algorithm based on global image information. Cluster Computing, 2019, 22(4): 8939–8951
3. Hossari M, Dev S, Nicholson M, Mccabe K, Nautiyal A, Conran C, et al. ADNet: a deep network for detecting adverts. In: Proceedings of CEUR Workshop. 2018, 45–53
4. Qian X, Tang G. Research on TV advertisement detection base on video shot. In: Proceedings of the 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization. 2012, 245–248
5. Zhang B, Xu B. Context-dependent audio-visual and temporal features fusion for TV commercial detection. In: Proceedings of International Symposium on Circuits and Systems. 2013, 5–8
6. Hannane R, Elboushaki A, Afdel K, Naghabhushan P, Javed M. An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram. International Journal of Multimedia Information Retrieval, 2016, 5(2): 89–104